Digital Open Science Index, Computer and Information Engineering Vol:14, No:11, 2020 waset.org/publication/10011595

Hybrid Structure Learning Approach for Assessing the Phosphate Laundries Impact

Emna Benmohamed, Hela Ltifi, Mounir Ben Ayed

***Abstract*—**Bayesian Network (BN) is one of the most efficient classification methods. It is widely used in several fields (i.e., medical diagnostics, risk analysis, bioinformatics research). The BN is defined as a probabilistic graphical model that represents a formalism for reasoning under uncertainty. This classification method has a high-performance rate in the extraction of new knowledge from data. The construction of this model consists of two phases for structure learning and parameter learning. For solving this problem, the K2 algorithm is one of the representative data-driven algorithms, which is based on score and search approach. In addition, the integration of the expert's knowledge in the structure learning process allows the obtainment of the highest accuracy. In this paper, we propose a hybrid approach combining the improvement of the K2 algorithm called K2 algorithm for Parents and Children search (K2PC) and the expert-driven method for learning the structure of BN. The evaluation of the experimental results, using the well-known benchmarks, proves that our K2PC algorithm has better performance in terms of correct structure detection. The real application of our model shows its efficiency in the analysis of the phosphate laundry effluents' impact on the watershed in the Gafsa area (southwestern Tunisia).

***Keywords*—**Classification, Bayesian network; structure learning, K2 algorithm, expert knowledge, surface water analysis.

I. Introduction

WATER scarcity is a real threat in our modern world. In fact, water war has been a contested topic [1] and researchers in this field assert that the world would attend a war over waters in the near future [2]. Citing as an example, the increasing of hydro political tensions among Mekong nations leads to war [3]. Therefore, protecting water resources from pollution and overuse has become an urgent priority. In this context, it seems very important to analyze the phosphate laundry effluents' impact on the watershed in Gafsa area (southwestern Tunisia) [4]. Data mining techniques used for classification is an open question in active research, especially for real and vital fields like water analysis.

BN is a probabilistic model representation in graphical mode which is constructed over a set of variables (random). BN is one of the most consistent formalisms for complex systems modeling [5]. In BN, the learning tasks can be subdivided into structure learning and parameter learning. Structure learning subtask specifies a set of the dependency relations between the random variables while the subtask of parameter learning allows the quantification of how strong are these dependencies. The structure learning phase produces a directed acyclic graph (DAG). Detecting the optimal structure is considered as an NP-hard problem due to the fact that the space of search is intractable [6]. In order to solve the above mentioned problem, data driven (or data based) method and expert driven BN structure learning method have been proposed [7].

In this paper, we present a hybrid approach based on the combination of automatic data analysis and expert knowledge based method for BN learning. In our contribution, we propose two improvements in order to increase the effectiveness of the classification technique: enhanced K2 algorithm called K2 algorithm for parents and child's detection (K2PC) and cognitive processes based structure learning amelioration. The proposed BN was experimented while assisting in the exploration of the water samples extracted from Gafsa zone (location shown in Fig. 8) [4].

Section II of this paper is devoted for presenting the interrelated notions before describing the main idea of our proposal. Section III presents the hybrid approach which is based on data oriented and expert oriented methods. Section IV presents an illustrative example. Section V details the experimental dataset and results. The last section is dedicated to the conclusion.

II. Preliminaries

Since we aim to propose a hybrid BN based on the K2 algorithm improvement and the knowledge and experiences of the expert, we have to present the interrelated principal notions in this section before highlighting the introduced ideas.

A. BNs

BNs are DAGs which enable efficient representation of the conditional probability distribution over a random variables set. The DAG consists of vertices (variables) and edges (dependency between the variables). The BN can be delineated as (G, P) a directed graph G and probability distributions P, where the graph is denoted as (N, E), in which N represents the nodes (or variables) and E are the edges. The nodes are denoted N = (N1, N2,…, Nn), each node (Ni) can have discrete or continuous values, the edges represent the dependency relationships between the variables. Accordingly, the joint probability distributions are calculated as the product of local conditional probability distributions. It is defined in:

$$P(N_1, N_2, ..., N_n) = \prod_{i=1}^{n} P(N_i \mid P_a(N_i))$$

Emna Benmohamed is with the University of Sfax, Tunisia (e-mail: emna.benmohamed.tn@ieee.org).

where N_i is a node and $P_a(N_i)$ is its parent.

The BN learning task over a set of given dataset's attributes consists of two learning phases, firstly of their qualitative component (topology or structure) and then of their quantitative component (parameters). To find the correct structure which better matches the dataset, three approaches have been proposed: the constraint-based, score-based and hybrid approaches. The score based approach contains the most widely used categories of algorithms like the K2 algorithm [8]-[10] that represents the topic of the following section.

B. K2 algorithm

Several efficient algorithms have been introduced for learning the structure of the BN like the K2 algorithm [12]. This latter is based on a greedy-heuristic search method for topology construction. In addition, the effectiveness of K2 algorithm depends mainly on the received order of variables. The pseudocode for K2 algorithm is indicated in Table I.

TABLE I
K2 ALGORITHM

Input: A set of n nodes, an ordering on the nodes, an upper bound u on the number of parents a node may have, and a database D containing m cases.
Ouput: The set of parents of each node.

```
for i:= 1 to n do
    $\pi i := \emptyset$;
    Pold := f(i, $\pi i$);
    OKToProceed := true;
    While OKToProceed and $|\pi i| < u$ do
        let z be the node in Pred(xi) - $\pi i$ that maximizes f(i, $\pi i \cup \{z\}$);
        Pnew := f(i, $\pi i \cup \{z\}$);
        if Pnew > Pold then
            Pold := Pnew;
            $\pi i := \pi i \cup \{z\}$;
        else OKToProceed := false;
    end while;
end for;
```

Actually, unless the perfect algorithm for the exact topology learning is proposed, the question "Who learns better Bayesian network structures?" remains unanswered [13].

C. Cognitive Process

User interaction with the visual representation, as a learning experience, consists of a set of cognitive activities, which are necessary for sense making and knowledge gaining [19], [15]. Furthermore, in order to reduce complexity, the main cognitive process is the data comprehension aiding in 'correct' mental image creation. Also, in cognition field, the activity of problem solving was considered as a high cognitive activity. In addition, Bloom's taxonomy of the cognitive domain (Bloom's taxonomy) [14] is widely used, well established and easily understood. Bloom introduced six levels of cognitive activities [19] as described in Table II.

TABLE II
BLOOM'S COGNITIVE ACTIVITIES

Level	ACTIVITY
1	Recognition
2	Comprehension
3	Application
4	Analysis
5	Synthesis
6	Evaluation

Recently, Ltifi et al. [15] presented a new cognitive process for extracted pattern analysis. This process consists of four activities which are perception, recognition, comprehension and reasoning as detailed in Table III.

TABLE III
COGNITIVE ACTIVITIES

Cognitive phases	Description
Perception	It presents the human ability to perceive visual elements that composes the visualization [16], [17]. The human perceives in parallel or sequentially these graphical elements which is considered as low cognitive activities.
Recognition	The comparison of visual elements with the memorized representation represents the recognition activity [18].
Comprehension	After the cognitive activity of visualizations recognition, the user understands the visualized elements. The trust of understandable elements contributes in novel knowledge generation [19].
Reasoning	This cognitive activity is based on the visual thinking of the user. This latter analyses the displayed information basing on the skills visual reasoning.

In the previous sections, we presented preliminary notions related to our proposal. We aim to create a semi-automatic classification method combining the improved BN and the expert experiences and knowledge for obtaining effective results.

III. PROPOSED APPROACH

To answer the above mentioned question, our proposed contribution is a hybrid approach combining data based method and expert based method for the creation of an effective classification technique. First, we propose an improvement of data oriented approach by introducing an amelioration of K2 algorithm for learning the structure of BN. Second, we suggest cognitive activities which are necessary for the produced model analysis.

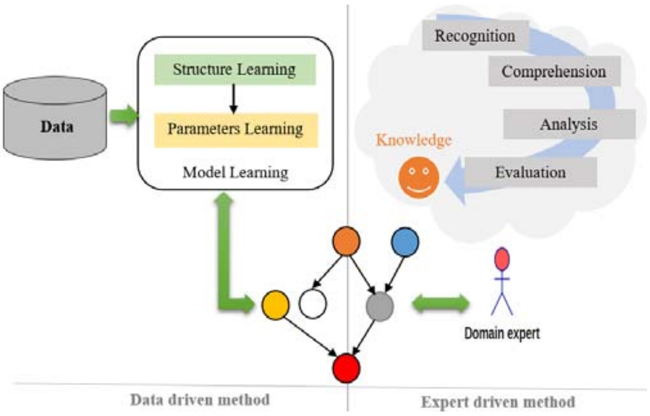

Fig. 1 BN learning based on data oriented method and expert oriented method

As shown in Fig. 1, the left side indicates the creation of the automatic method based BN model including the two phases of BN learning from data. For structure learning, we introduce an improved K2 algorithm with extended space of research. Then the probability distributions can be calculated based on the generated BN topology. Therefore, as visualized in the figure, the resulted oriented acyclic graph that shows the variables dependency. After perceiving the graph, the expert executes a set of cognitive activities for new knowledge generation. In parallel, he/she is able to ameliorate the topology as shown using the bidirectional arrow. These human activities represent the expert driven method used for the model building. The two methods will be detailed in the following sections.

A. K2PC Algorithm for the BN Structure Learning

The main idea of K2 is to search the node that maximizes the calculated score. Thus, the preceding nodes set is considered as the effective search space of parents for each node. Accordingly, extending the search space and including the children search space represents the principle of our improvement. For each node, we subdivided the search space according to the given order on the parents' search space (set of preceding nodes) and the search space of children (set of succeeding nodes). Hence, the proposed name K2PC that denotes the K2 algorithm for parents and children search.

According to the given order, the K2PC algorithm tests the entered nodes sequentially. In fact, the proposed idea can be recapitulated through the following description:

For each node Ni,

Step1. We define the parents set of Xi (Pa(Xi)) as empty set.

Step2. Then, the K2PC algorithm calculates the local score.

Step3. We select the node that better increases the local score from the search-space of Xi parents. This node will be added to the Pa(Xi).

Step4. We continue the execution of previous steps (1, 2 and 3) until no node is able to increase the score or the maximum number of parents is reached.

Step5. We define the set of Xi children (Ch(Xi)) as empty set.

Step6. We calculate the local score of this node.

Step7. We select the node that further decreases the score from the set of succeeding nodes.

Step8. Previous steps for children detection continue the execution until there is not a node allowing the decrease in the score.

Step9. If the selected-score's absolute value is less than the best-parent-score absolute value, the node will be added as a child of Xi.

The proposed algorithm is a specific method of selecting parents and children for the Xi node. In particular, we extended the search space for parents and children detection by respecting the given order. The generated graph is acyclic and oriented in order to visualize the dependency relationships between variables. Accordingly, the perceiver can analyze the K2PC algorithm's output by executing a set of cognitive activities as explained in the following section.

B. Expert Knowledge Driven Method for the BN Structure Learning

Our second approach of the pattern analysis using the cognitive process is based on the expert's knowledge and experiences. Our aim is to explore the generated graph and probabilities by discovering significant information. The analysis of visualized dependency relationships between variables allows the expert to gain new knowledge. During the analysis process, the expert starts by recognizing the represented variables and links. The strength of an edge relating two nodes is measured by conditional probability. Indeed, he/she compares the obtained results with memorized information among his/her experiences. Then the expert perceives again and again the different visual representations (nodes and arcs) and comprehends the relationships between variables by employing the human expertise for patterns analysis. Finally, the produced patterns can be evaluated and the expert may ameliorate generated results (adding, inversing and deleting visual elements). In addition, he/she determines the useful relationships from the given representation of the situation. Accordingly, the expert gains new knowledge and produced BN model will be improved.

IV. Illustrative Example

In this section, we present an illustrative example that explains clearly our contribution. We used one of the well-known benchmarks for BN learning named ASIA. This latter consists of 8 nodes linked by 8 edges. For instance, we give the order of nodes (1, 2, 3, 4, 5, 6, 7, 8) as input for our algorithm. By applying the K2PC algorithm for the node 5 situated in the position number 5, the parents-search-space is colored in green. However, the nodes 6, 7 and 8 which are in orange color represent the nodes that can be added to the children set of the node 5.

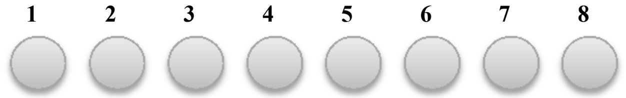

Fig. 2 Asia network in natural order

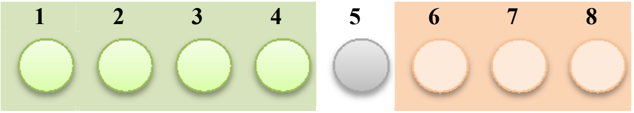

Fig. 3 Subdividing the search space for parents and children search

After executing the K2PC coded in MATLAB on a machine which is Intel Core i5 with 8GB RAM running on Windows 10 operating system using BNT project, the produced graph is represented as in Fig. 4.

Our algorithm is effective for the BN structure learning, as shown in Fig. 4; the resulted structure by applying the improved K2 algorithm on Asia network is similar to the original structure. In [20] we detailed the produced results and we evaluated the K2PC algorithm's effectiveness comparing to the previous works. In Figs. 5 and 6, we will present the comparison of structure difference between the original topology and the resulted one. In order to demonstrate the effectiveness of our algorithm, we use the accuracy factor that is defined in [11]:

$$Accuracy = \frac{correctedges}{correctedges + erroredges} \quad (1)$$

We determine the accuracy values for three well-known benchmarks which are Cancer, Asia and Alarm (1000, 2000, 5000 and 10000 sets of data). The produced results by the improved K2 algorithm are illustrated in Fig. 5 where the terms AE, DE, RE and CE designate respectively the number of added, deleted, reversed and correct edges compared to the original topology which is used to prove the correctness of the learned structure. For instance, the proposed algorithm allows obtaining 42 correct edges, 0 reversed edges, 4 deleted edges and one added edges (more details are in [20]).

The proposed K2PC algorithm represents an effective method for learning the structure of BN comparing to some algorithms as explained in [20]. In our case the expert in medical field who uses this data recognizes the different variables and links, comprehends the dependency relationships, then analyses the gained information and finally evaluates the extracted knowledge for new knowledge production.

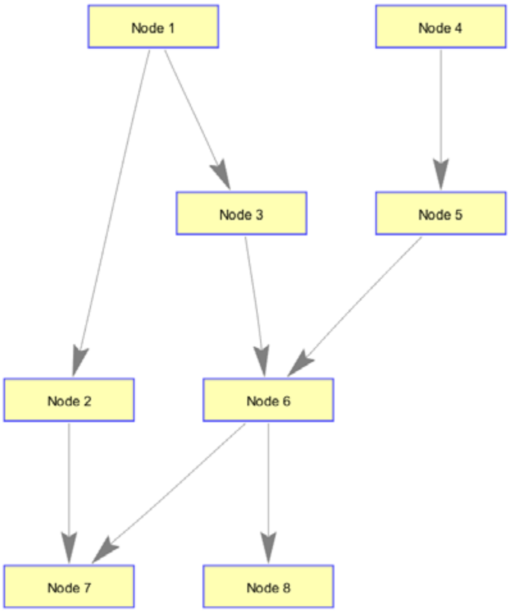

Fig. 4 The learned structure by K2PC algorithm

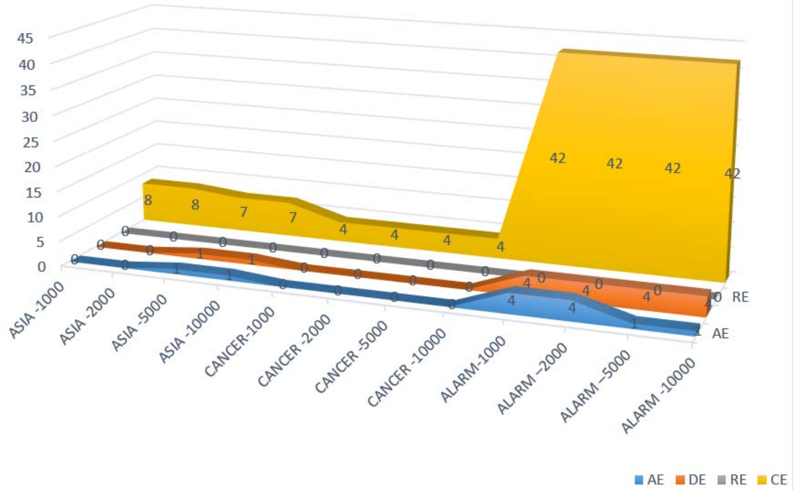

Fig. 5 Results given by K2PC algorithm

V. Experiments and Results

This section is dedicated to the presentation of geologic dataset and the evaluation measures used, then to display the produced results. Finally, a discussion section is conducted for the evaluation of the correctness of the proposed combination of data driven method and expert oriented method.

A. The Used Dataset

As mentioned above, the preservation of water resources is a vital task. In this context, [4] extracted several samples of water from Gafsa that is an industrial zone known by the extraction of phosphate. The natural phosphate processing in various laundries allows the increasing of the P_2O_5 and the elimination of undesired fraction. The effluents of laundries contain different contaminants which are discharged to the receiving water body. Consequently, the collection of samples from different locations of the phosphatic area resulted in 107 samples where 80 are for the training dataset and 27 reserved for the test dataset. These 11 variables are*: instance number 'num', chemical variables which are sec residu (g/l), Ca^{2+} (mg/l), Na^{+} (mg/l), SO_4^{2-} (mg/l), HCO_3^{-} (mg/l), Cd (mg/l), Fluor (mg/l), salinity (g/l), NO_3 (mg/l) and class.*

B. Learning the BN Model

The application of BN models in real dataset represents a challenge for the researchers in the classification field due to the necessity of constantly regulating the appropriate model. This calibration lies in the identification of the correct graph and of the most appropriate parameters. In fact, the estimation of BN model consists of two phases which are the BN structure learning and the learning of BN parameters. The first subtask is achieved using the proposed K2PC algorithm (described in Section III). Then for learning the parameters, we used the EM algorithm [21] presuming a fixed topology.

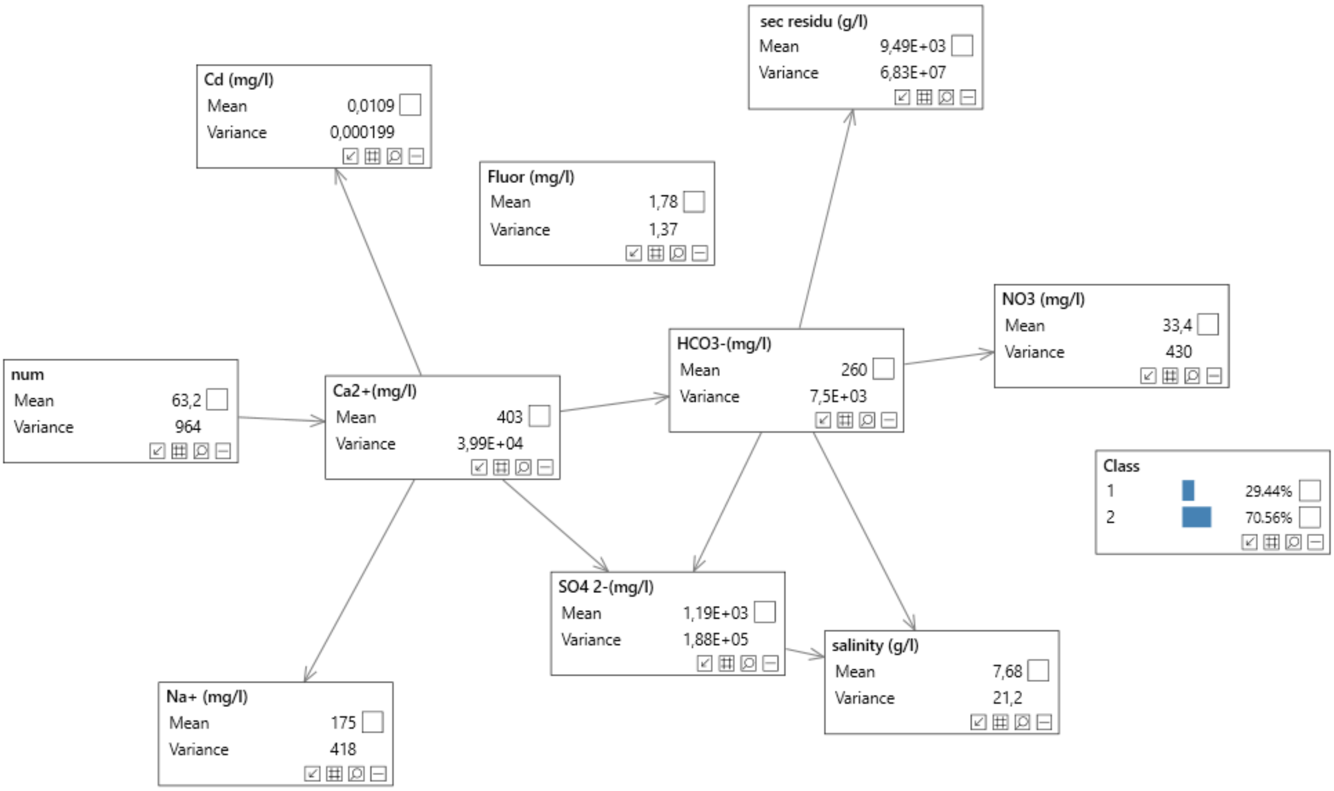

Fig. 6 The resulted BN model for the geologic dataset

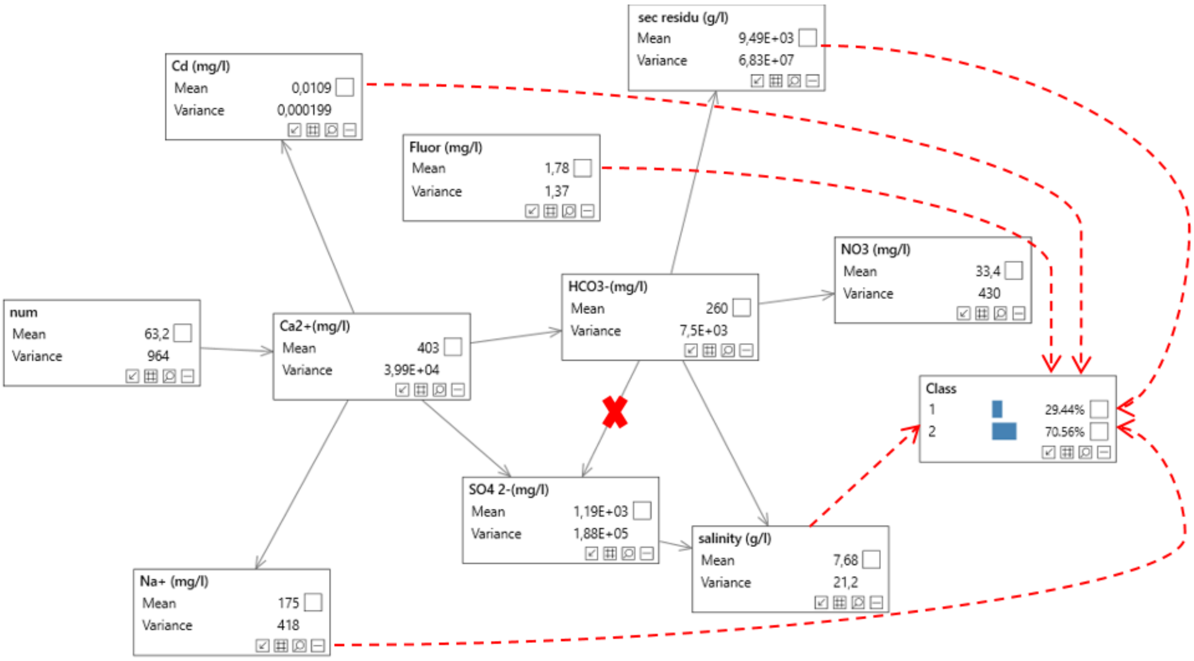

Fig. 7 The BN model for the geologic dataset

The obtained model represents the result of the execution of the two algorithms (the K2PC algorithm for structure learning and EM algorithm for the learning of the parameters). The model reported in Fig. 6 procures the dependency relationships between geologic variables for the water samples analysis as described in Section *A*. The expert discovers that there are missing edges which are important in the model calibration. For instance, the node Cd must be related to the node Class. The values of the Cd variable are very important to classify the instance (If the concentration of Cadmium in water samples is below 3 µg/l, the class is 1 that means that is a good sample (1), else the class is not good (2)). In addition, he/she proposes a list of modifications leading the obtainment of the appropriate BN model as shown in Fig. 7.

In this work, BN model was built using a hybrid method based on improved K2 algorithm and expert knowledge for geologic dataset analysis. Aiming to facilitate the user's tasks, we propose an automatic and visual data analysis for water examples' classification.

These results have revealed the advantages of integrating the expert's knowledge and experiences in the BN model learning. As shown, the quality of gained topology is useful for examples analysis in order to demonstrate the phosphate laundry effluents' impact on the watershed in Gafsa area (southwestern Tunisia).

VI. Conclusion

In this paper, we proposed a hybrid approach for BN model learning based on data driven method and expert knowledge oriented method. The proposal has been applied for real data analysis in a vital field. As shown in the experimental results, both automatic and human methods are extremely useful for the creation of the appropriate BN model. This paper represents an extension of improved approach for learning the structure of BN. The obtained results in terms of internal and external validity indexes show the efficiency of the improvement. Effectively, a future work can be proposed to demonstrate the importance of the gained model for the classification of the extracted examples.

Appendix

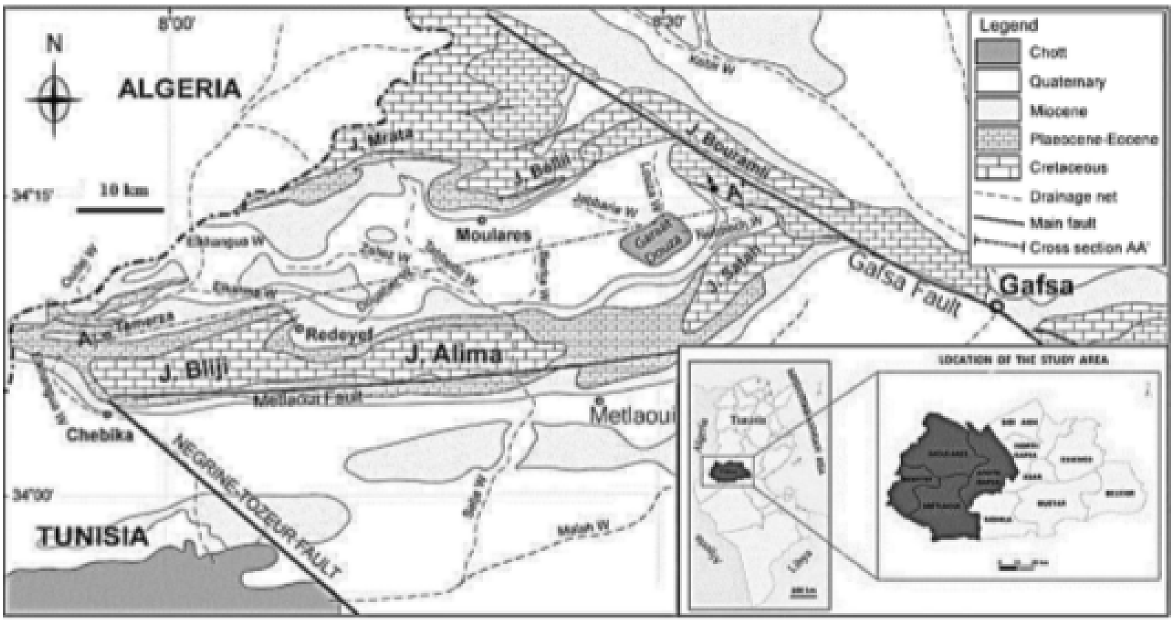

Fig. 8 Location of the study area in the simplified geologic map [4]

References

[1] Z. Wang, Why Does the Water War Thesis -Prevail? International Co-operation and Development, European Commission, 2013.

[2] S. Pradhan, "Water War Thesis: A Myth or A Reality? ". International Journal of Arts, Humanities and Social Science, vol 2, no. 1, pp. 12-15, 2017.

[3] Pearse-Smith, S. W. "Water war'in the Mekong Basin? ". Asia Pacific Viewpoint,vol. 53, no. 2, pp. 147-162, 2012.

[4] S. Marzougui, A. Sdiri, & F. Rekhiss. "Heavy metals' mobility from phosphate washing effluents discharged in the Gafsa area (southwestern Tunisia) ". Arabian Journal of Geosciences, vol. 9, no. 12, 599, 2016.

[5] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, G. Condous, "Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression", Human Reproduction, vol. 21, no. 7, pp. 1824–1831, 2006, https://doi.org/10.1093/humrep/del083

[6] W. Buntine, "A guide to the literature on learning probabilistic networks from data," IEEE Trans. Knowl. Data Eng., vol. 8, no. 2, pp. 195–210, 1996.

[7] H. S. Sousa, F. Prieto-Castrillo, J. C. Matos, J. M. Branco, & P. B. Lourenço, "Combination of expert decision and learned based Bayesian Networks for multi-scale mechanical analysis of timber elements". Expert Systems with Applications, vol. 93, pp. 156-168, 2018.

[8] V. R. Tabar,, F. Eskandari, S. Salimi, et al. " Finding a set of candidate parents using dependency criterion for the K2 algorithm". Pattern Recognition Letters, vol. 111, pp. 23-29, 2018.

[9] H. Amirkhani, M. Rahmati, P. J. Lucas, & A. Hommersom, "Exploiting experts' knowledge for structure learning of Bayesian networks". IEEE transactions on pattern analysis and machine intelligence, vol 39, no 11, pp. 2154-2170, 2016

[10] S. Aouay, S. Jamoussi,, & Y. B. Ayed, " Particle swarm optimization based method for Bayesian Network structure learning". In 5th International Conference on Modeling, Simulation and Applied

Optimization (ICMSAO) , pp. 1-6. IEEE, 2013.

[11] M. Scutari, C. E. Graafland & J. M. Gutiérrez,“ Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms“. International Journal of Approximate Reasoning, vol. 115, pp. 235-253, 2019.

[12] G.F. Cooper, E. Herskovits, “A Bayesian method for the induction of probabilistic networks form data “, Mach. Learn. Vol. 9, pp. 309–347, 1992.

[13] L. Huang, G. Cai, H. Yuan, & J. Chen, “A hybrid approach for identifying the structure of a Bayesian network model“. Expert Systems with Applications, vol. 131, 308-320, 2019.

[14] B. S. Bloom, “Taxonomy of educational objectives: The classification of educational goals“, Cognitive domain, 1956.

[15] H. Ltifi, E. Benmohamed, C. Kolski, M. Ben Ayed M, “Adapted visual analytics process for intelligent decision-making: Application in a medical context“, *International Journal of Information Technology & Decision Making 2019,* accepted paper.

[16] D. Pineo, and C. Ware, “Data visualization optimization via computational modeling of perception“, IEEE Trans. on Visualization and Computer Graphics, vol. 18, no. 2, pp. 309-320, 2012.

[17] A. Pineo, T-D Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman, “Interactive information visualization to explore and query electronic health records”, Found Trends Hum–Comput Interact, vol. 5, no. 3, pp. 207–298, 2013.

[18] J. Zheng, Z. Jiang, R. Chellappa, “Cross-view action recognition via transferable dictionary learning, ” IEEE Trans. Image Process, vol. 25, no. 6, , pp. 2542-2556, 2016.

[19] E. Benmohamed, H. Ltifi, H., M. B. Ayed, ”Using Bloom's taxonomy to enhance interactive concentric circles representation”. In 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.

[20] E. Benmohamed. H. Ltifi, M. Benayed “A Novel Bayesian Network Structure Learning Algorithm: Best Parents-Children,” in proceeding. IEEE *ISKE, t*he 14th International Conference on Intelligent Systems and Knowledge Engineering, *2019*.

[21] K. Masmoudi, L. Abid, &A. Masmoudi, “Credit risk modeling using Bayesian network with a latent variable,” vol. 127, Expert Systems with Applications, pp. 157-166, 2019.